# DUOS:

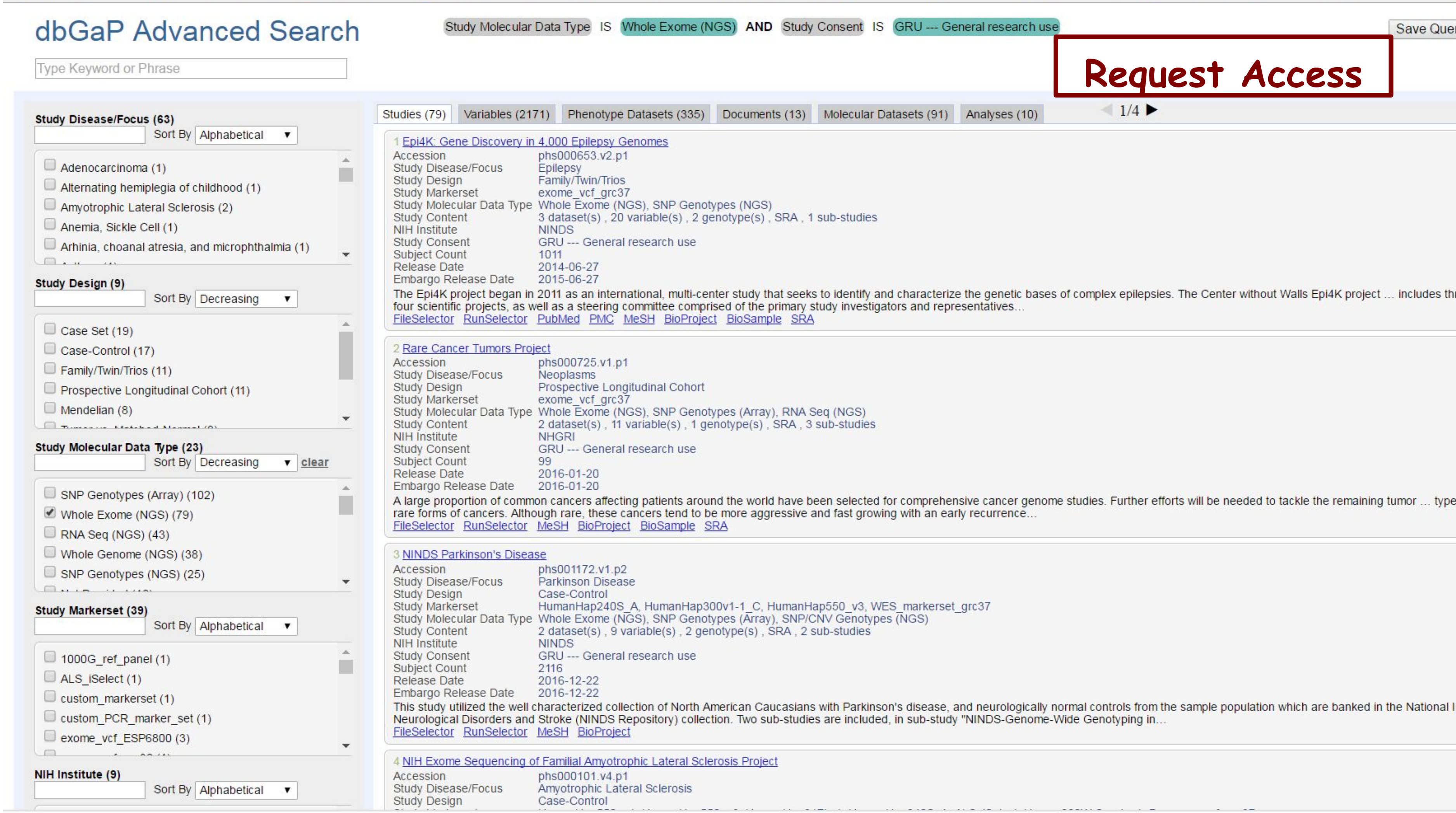
# *A Structured Approach to Genomics’ Data Use Oversight*

https://duos.broadinstitute.org

BROAD INSTITUTE

*Moran Cabili*
*nmcabili@broadinstitute.org*

# If Time From Discovery to Access Was Instant...



dbGaP Advanced Search

Study Molecular Data Type IS Whole Exome (NGS) AND Study Consent IS GRU --- General research use

Save Quer

Request Access

Type Keyword or Phrase

**Study Disease/Focus (63)**
Sort By Alphabetical ▼

- ☐ Adenocarcinoma (1)
- ☐ Alternating hemiplegia of childhood (1)
- ☐ Amyotrophic Lateral Sclerosis (2)
- ☐ Anemia, Sickle Cell (1)
- ☐ Arhinia, choanal atresia, and microphthalmia (1)

**Study Design (9)**
Sort By Decreasing ▼

- ☐ Case Set (19)
- ☐ Case-Control (17)
- ☐ Family/Twin/Trios (11)
- ☐ Prospective Longitudinal Cohort (11)
- ☐ Mendelian (8)

**Study Molecular Data Type (23)**
Sort By Decreasing ▼ clear

- ☐ SNP Genotypes (Array) (102)
- ☑ Whole Exome (NGS) (79)
- ☐ RNA Seq (NGS) (43)
- ☐ Whole Genome (NGS) (38)
- ☐ SNP Genotypes (NGS) (25)

**Study Markerset (39)**
Sort By Alphabetical ▼

- ☐ 1000G_ref_panel (1)
- ☐ ALS_iSelect (1)
- ☐ custom_markerset (1)
- ☐ custom_PCR_marker_set (1)
- ☐ exome_vcf_ESP6800 (3)

**NIH Institute (9)**
Sort By Alphabetical ▼

Studies (79) | Variables (2171) | Phenotype Datasets (335) | Documents (13) | Molecular Datasets (91) | Analyses (10)

◀ 1/4 ▶

1 Epi4K: Gene Discovery in 4,000 Epilepsy Genomes

| | |
|---|---|
| Accession | phs000653.v2.p1 |
| Study Disease/Focus | Epilepsy |
| Study Design | Family/Twin/Trios |
| Study Markerset | exome_vcf_grc37 |
| Study Molecular Data Type | Whole Exome (NGS), SNP Genotypes (NGS) |
| Study Content | 3 dataset(s) , 20 variable(s) , 2 genotype(s) , SRA , 1 sub-studies |
| NIH Institute | NINDS |
| Study Consent | GRU --- General research use |
| Subject Count | 1011 |
| Release Date | 2014-06-27 |
| Embargo Release Date | 2015-06-27 |

The Epi4K project began in 2011 as an international, multi-center study that seeks to identify and characterize the genetic bases of complex epilepsies. The Center without Walls Epi4K project ... includes thr four scientific projects, as well as a steering committee comprised of the primary study investigators and representatives...

FileSelector RunSelector PubMed PMC MeSH BioProject BioSample SRA

2 Rare Cancer Tumors Project

| | |
|---|---|
| Accession | phs000725.v1.p1 |
| Study Disease/Focus | Neoplasms |
| Study Design | Prospective Longitudinal Cohort |
| Study Markerset | exome_vcf_grc37 |
| Study Molecular Data Type | Whole Exome (NGS), SNP Genotypes (Array), RNA Seq (NGS) |
| Study Content | 2 dataset(s) , 11 variable(s) , 1 genotype(s) , SRA , 3 sub-studies |
| NIH Institute | NHGRI |
| Study Consent | GRU --- General research use |
| Subject Count | 99 |
| Release Date | 2016-01-20 |
| Embargo Release Date | 2016-01-20 |

A large proportion of common cancers affecting patients around the world have been selected for comprehensive cancer genome studies. Further efforts will be needed to tackle the remaining tumor ... type rare forms of cancers. Although rare, these cancers tend to be more aggressive and fast growing with an early recurrence...

FileSelector RunSelector MeSH BioProject BioSample SRA

3 NINDS Parkinson's Disease

| | |
|---|---|
| Accession | phs001172.v1.p2 |
| Study Disease/Focus | Parkinson Disease |
| Study Design | Case-Control |
| Study Markerset | HumanHap240S_A, HumanHap300v1-1_C, HumanHap550_v3, WES_markerset_grc37 |
| Study Molecular Data Type | Whole Exome (NGS), SNP Genotypes (Array), SNP/CNV Genotypes (NGS) |
| Study Content | 2 dataset(s) , 9 variable(s) , 2 genotype(s) , SRA , 2 sub-studies |
| NIH Institute | NINDS |
| Study Consent | GRU --- General research use |
| Subject Count | 2116 |
| Release Date | 2016-12-22 |
| Embargo Release Date | 2016-12-22 |

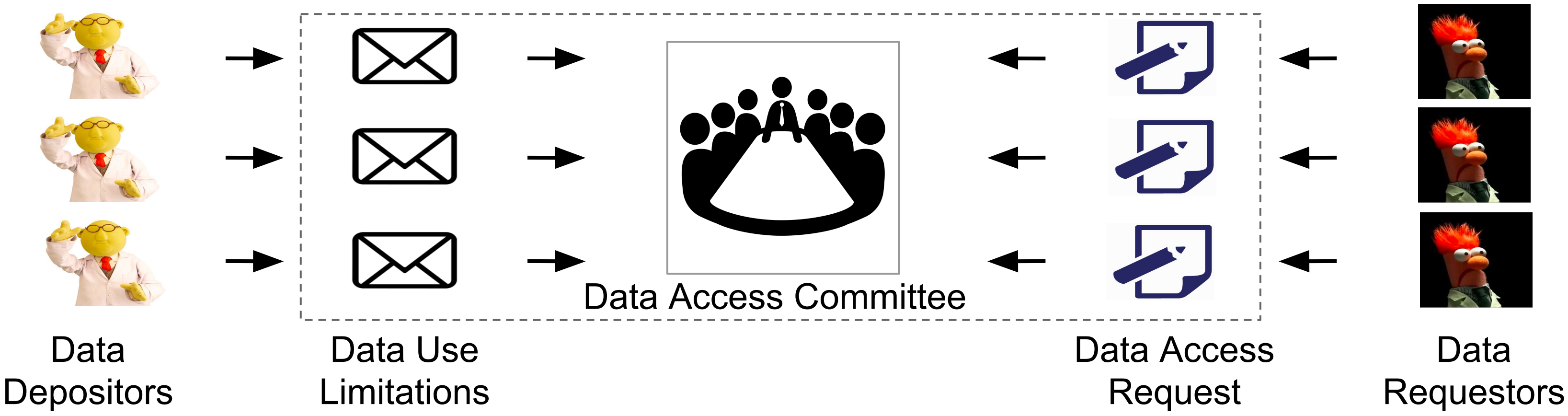
This study utilized the well characterized collection of North American Caucasians with Parkinson's disease, and neurologically normal controls from the sample population which are banked in the National I Neurological Disorders and Stroke (NINDS Repository) collection. Two sub-studies are included, in sub-study "NINDS-Genome-Wide Genotyping in...

FileSelector RunSelector MeSH BioProject

4 NIH Exome Sequencing of Familial Amyotrophic Lateral Sclerosis Project

| | |
|---|---|
| Accession | phs000101.v4.p1 |
| Study Disease/Focus | Amyotrophic Lateral Sclerosis |
| Study Design | Case-Control |

# Our Current Protocol for Data Access



Data Access Committee

Data Depositors

Data Use Limitations

Data Access Request

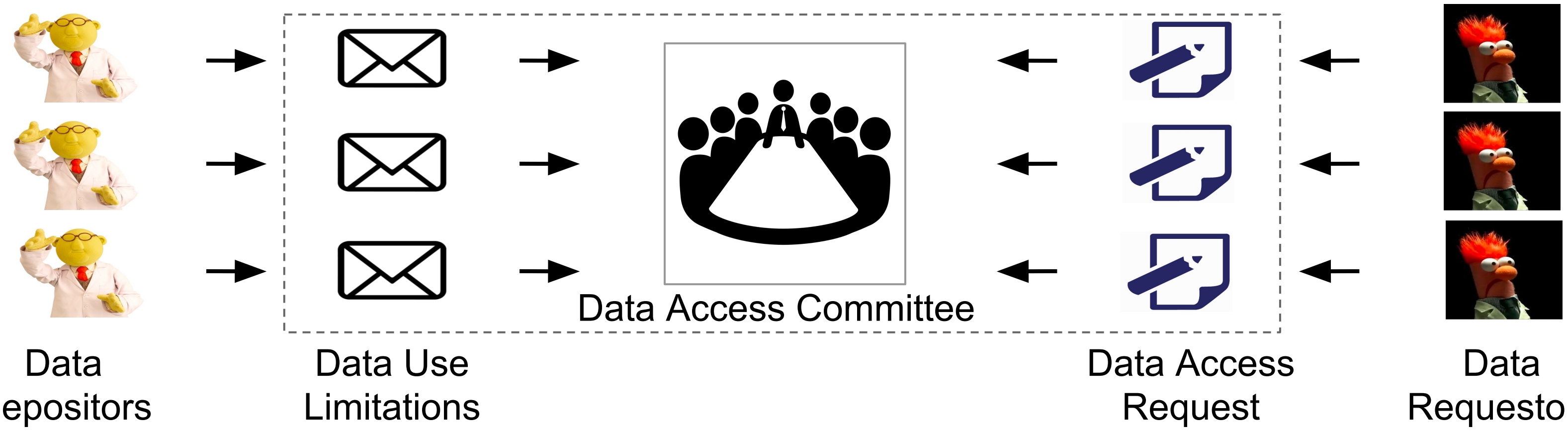
Data Requestors

***No!***

This data is available for cancer research in a non-profit setting.

I am studying Breast cancer at a company.

# Our Current Protocol for Data Access



Data Access Committee

Data Depositors

Data Use Limitations

Data Access Request

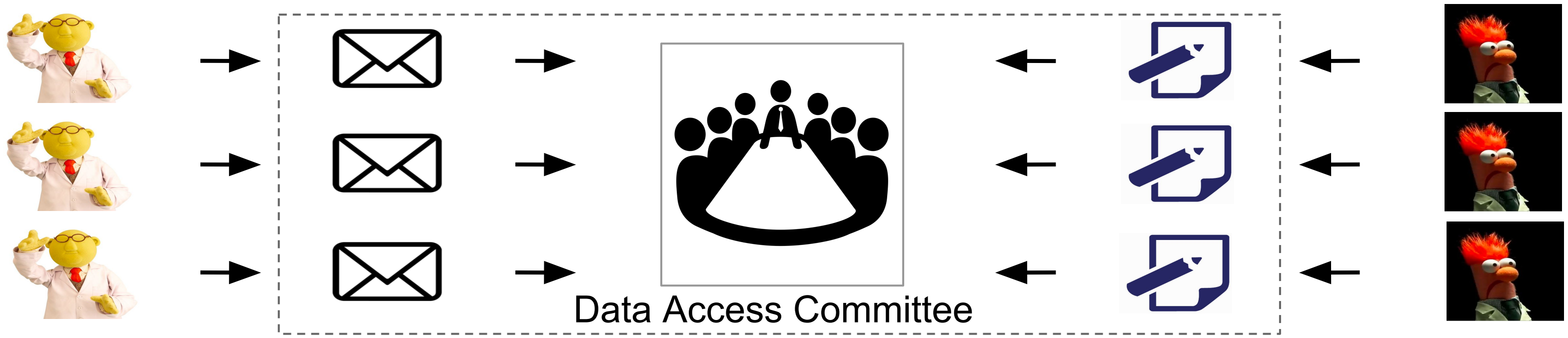
Data Requestors

***Yes!***

This data is available for cancer research in a non-profit setting.

I am studying Breast cancer at a non-profit.

# Our Current Protocol for Data Access



Data Access Committee

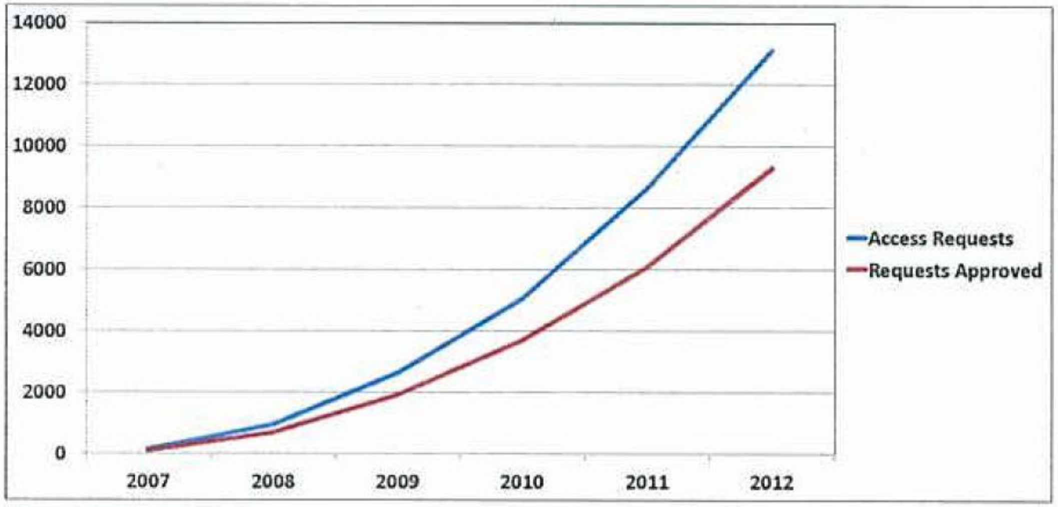
Data Depositors

Data Use Limitations

Data Access Request

Data Requestors

*Cumulative Number of Data Access Requests Submitted and Data Access Requests Approved, April 2007- November 2012*

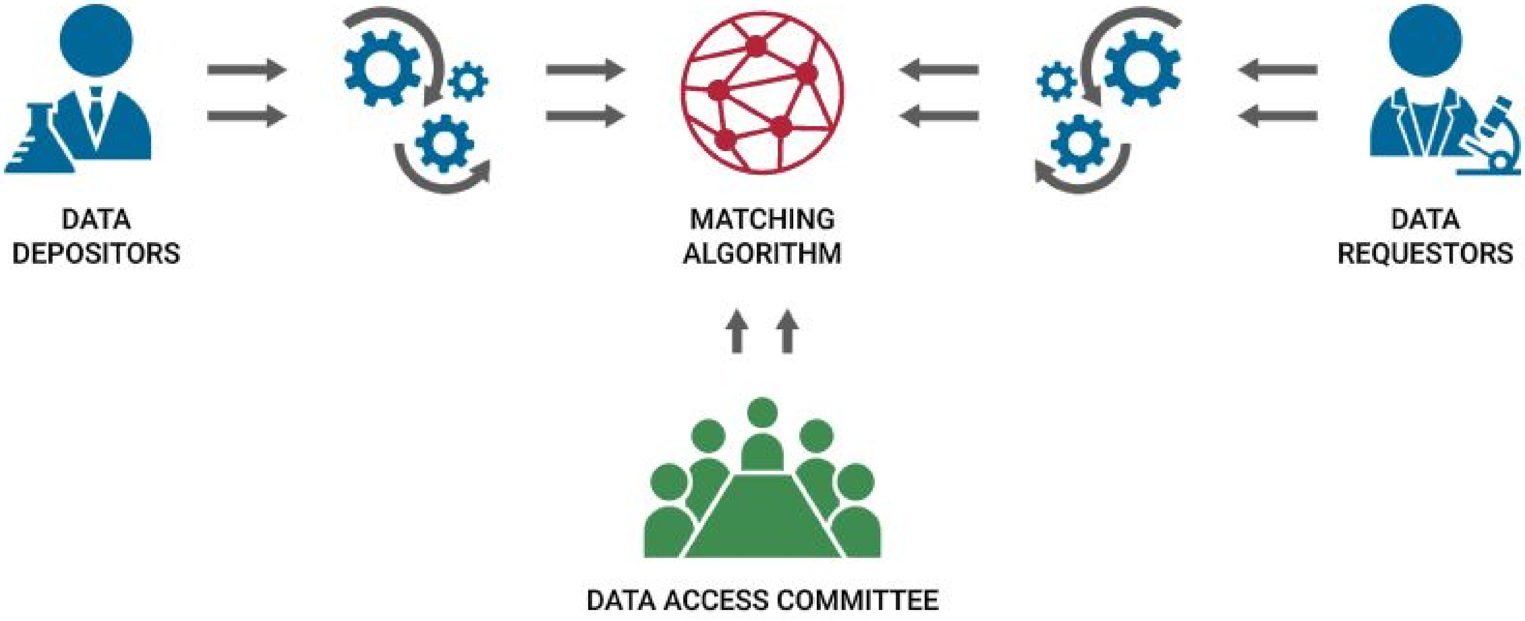


*Partial data for 2013 not shown*

***Scales Poorly!!***
***O($N^2$)***

dbGaP at PRIMr 2013

# DUOS



***What is DUOS?***

https://duos.broadinstitute.org/#/home

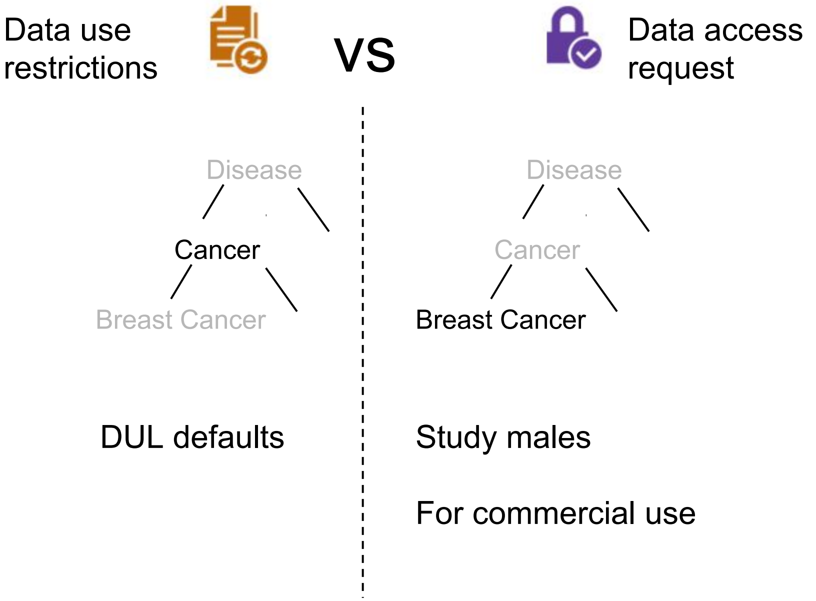
- Interfaces to transform data use restrictions and data access requests to machine-readable code (ADA-M & Consent Codes)
- A matching algorithm that checks if data access requests are compatible with data use restrictions
- Interfaces for the Data Access Committee to adjudicate whether structuring and matching has been done appropriately

# DUOS: Matching Algorithm

## Uses a Web Ontology Language (OWL) reasoner

### Consent Category

| | |
|---|---|
| Restricted to the study of Disease (ontology ID) | DS-[XX](CC)<br>**or** HMB(CC) |
| Commercial use is prohibited | NPU |
| General methods development prohibited | NMDS |
| Population structure or normal variation studies prohibited | NPNV |
| Samples use as controls outside the specified restrictions is prohibited | *NCTRL* |
| limited to gender (M/F/none) | RS-[GENDER] |
| limited to pediatric research | RS-[PEDIATRIC] |
| Sensitive/restricted population; Date | MAN-REV |


Data use restrictions

VS

Data access request

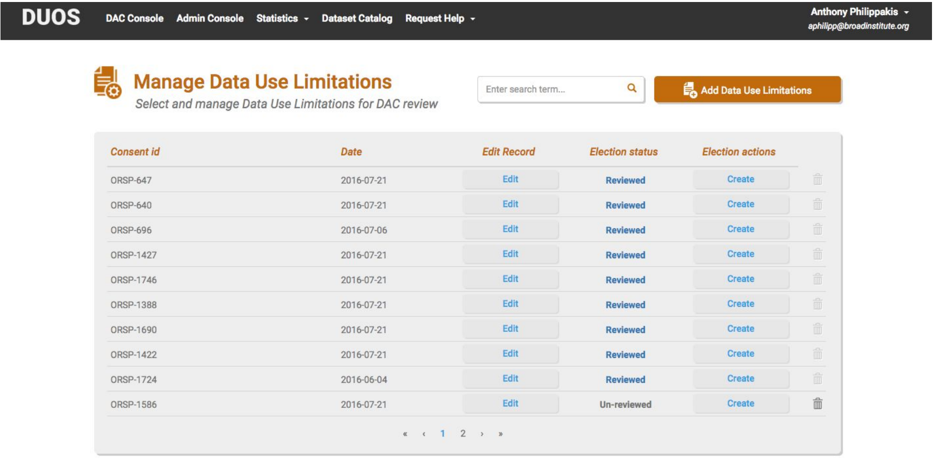
DUL defaults

Study males

For commercial use

Matching Algorithm
Answer: ***Yes!***

# DUOS: DAC Interfaces



DUOS DAC Console Admin Console Statistics ▾ Dataset Catalog Request Help ▾ Anthony Philippakis ▾ aphilipp@broadinstitute.org

**Manage Data Use Limitations**

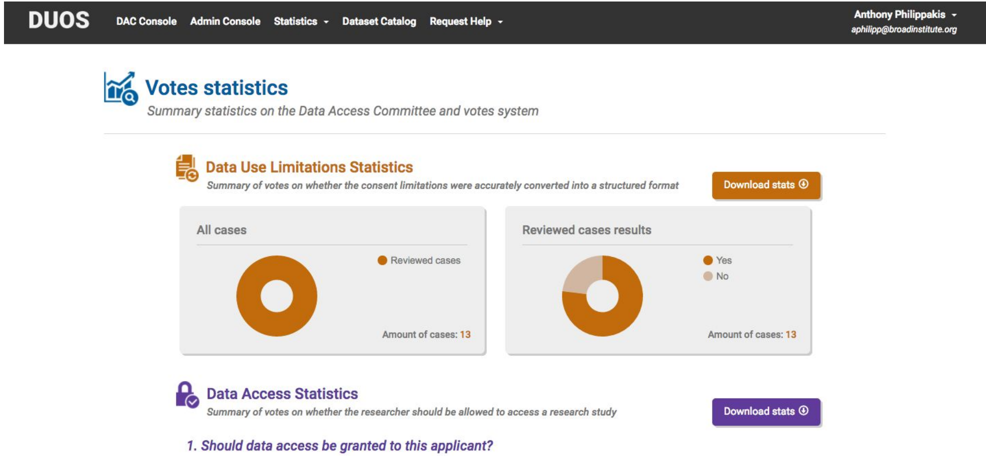
*Select and manage Data Use Limitations for DAC review*

Enter search term... | Add Data Use Limitations

| Consent id | Date | Edit Record | Election status | Election actions |
|---|---|---|---|---|
| ORSP-647 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-640 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-696 | 2016-07-06 | Edit | Reviewed | Create |
| ORSP-1427 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-1746 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-1388 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-1690 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-1422 | 2016-07-21 | Edit | Reviewed | Create |
| ORSP-1724 | 2016-06-04 | Edit | Reviewed | Create |
| ORSP-1586 | 2016-07-21 | Edit | Un-reviewed | Create |

« ‹ 1 2 › »

Structure & Manage Data Use Letters



DUOS DAC Console Admin Console Statistics ▾ Dataset Catalog Request Help ▾ Anthony Philippakis ▾ aphilipp@broadinstitute.org

**Votes statistics**

*Summary statistics on the Data Access Committee and votes system*

**Data Use Limitations Statistics**

*Summary of votes on whether the consent limitations were accurately converted into a structured format*

Download stats

All cases

● Reviewed cases

Amount of cases: 13

Reviewed cases results

● Yes
● No

Amount of cases: 13

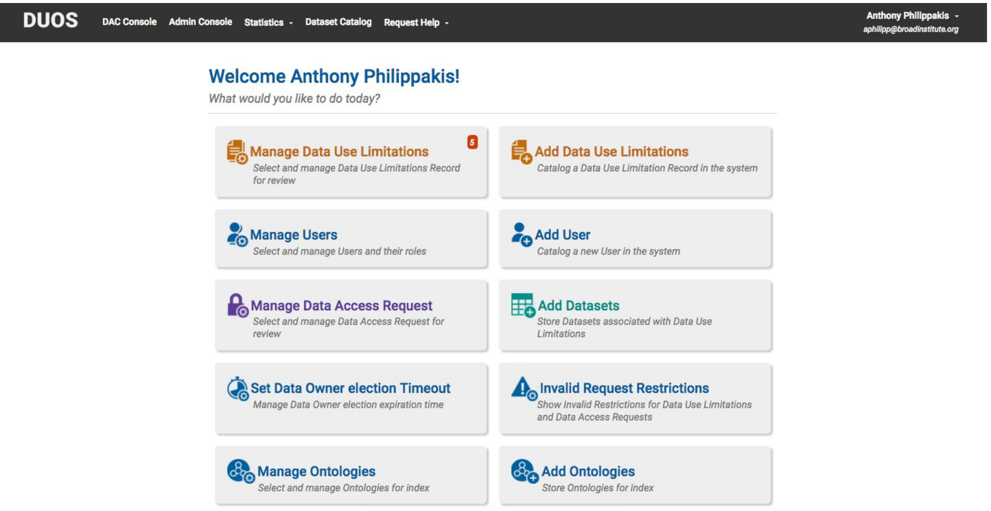
**Data Access Statistics**

*Summary of votes on whether the researcher should be allowed to access a research study*

Download stats

*1. Should data access be granted to this applicant?*

Approval of Structured Data Use



DUOS DAC Console Admin Console Statistics ▾ Dataset Catalog Request Help ▾ Anthony Philippakis ▾ aphilipp@broadinstitute.org

**Welcome Anthony Philippakis!**

*What would you like to do today?*

**Manage Data Use Limitations** 5
*Select and manage Data Use Limitations Record for review*

**Add Data Use Limitations**
*Catalog a Data Use Limitation Record in the system*

**Manage Users**
Select and manage Users and their roles

**Add User**
*Catalog a new User in the system*

**Manage Data Access Request**
*Select and manage Data Access Request for review*

**Add Datasets**
*Store Datasets associated with Data Use Limitations*

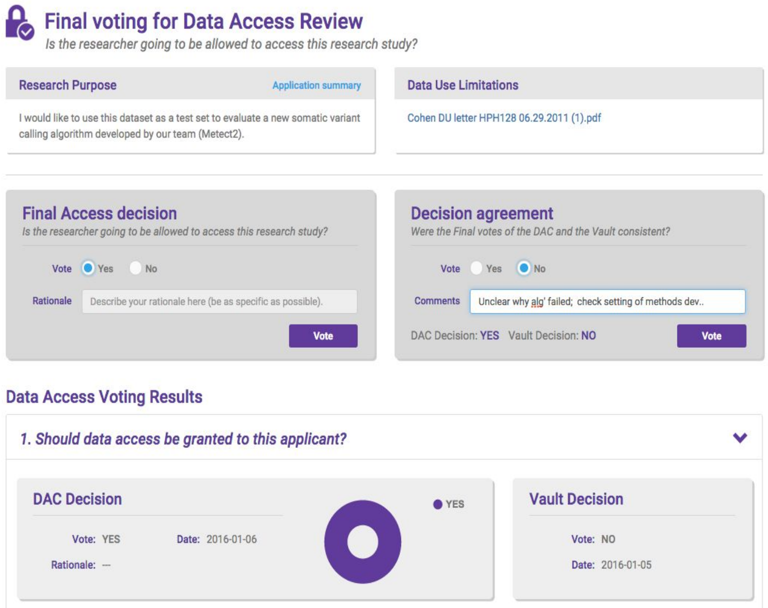
**Set Data Owner election Timeout**
*Manage Data Owner election expiration time*

**Invalid Request Restrictions**
*Show Invalid Restrictions for Data Use Limitations and Data Access Requests*

**Manage Ontologies**
*Select and manage Ontologies for index*

**Add Ontologies**
*Store Ontologies for index*

Administrative Console



**Final voting for Data Access Review**

*Is the researcher going to be allowed to access this research study?*

**Research Purpose** Application summary

I would like to use this dataset as a test set to evaluate a new somatic variant calling algorithm developed by our team (Metect2).

**Data Use Limitations**

Cohen DU letter HPH128 06.29.2011 (1).pdf

**Final Access decision**

*Is the researcher going to be allowed to access this research study?*

Vote ● Yes ○ No

Rationale: Describe your rationale here (be as specific as possible).

Vote

**Decision agreement**

*Were the Final votes of the DAC and the Vault consistent?*

Vote ○ Yes ● No

Comments: Unclear why alg' failed; check setting of methods dev..

DAC Decision: YES Vault Decision: NO

Vote

**Data Access Voting Results**

*1. Should data access be granted to this applicant?*

**DAC Decision**

Vote: YES Date: 2016-01-06

Rationale: ---

● YES

**Vault Decision**

Vote: NO

Date: 2016-01-05

DAC Members Approve Access

# Validation of DUOS

**Claim:** Data Use Can Be Structured

**Test:** Run a trial!

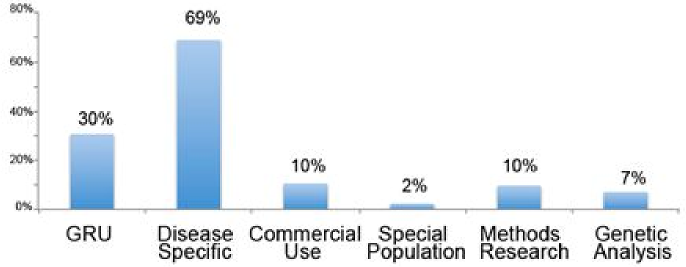


**Figure 2:** Data use restrictions can be structured by using 5 main categories (non-mutually exclusive; N=125)

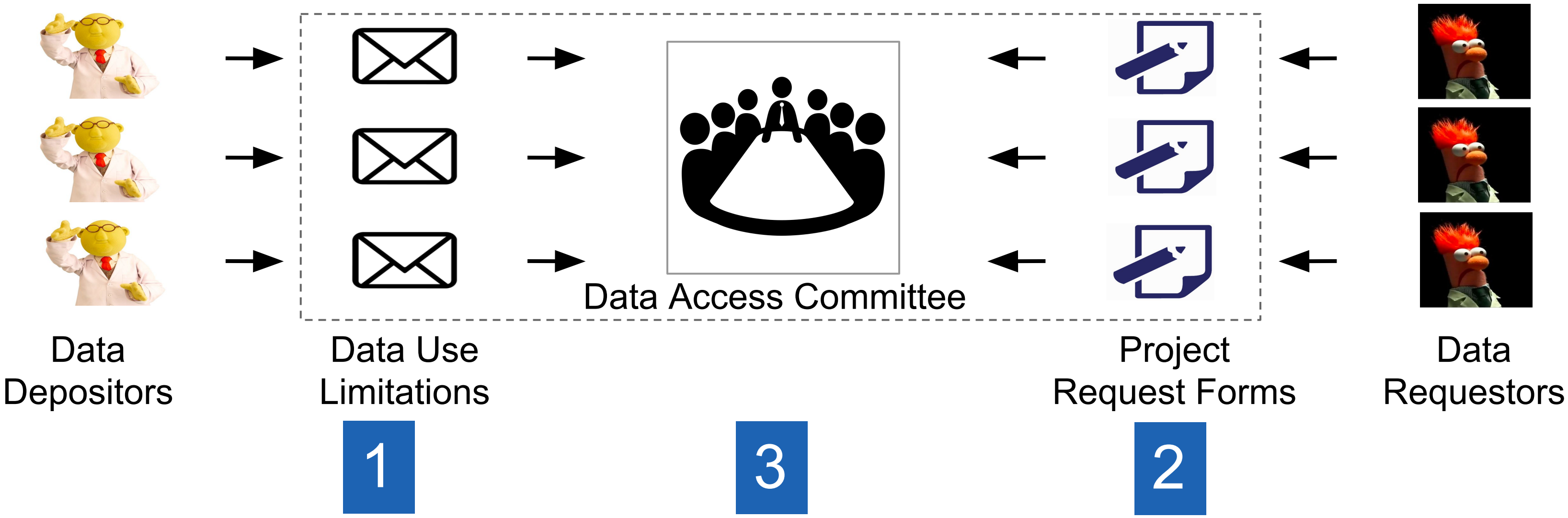
**Data Access Committee**
Pearl O'Rourke (Partners)
Laura Rodriguez (NIH)
John Wilbanks (Sage)
Stacey Donnelly (Broad)
Anthony Philippakis (Broad)

**Diseases:** Diabetes research only, Breast cancer research only, etc
**Commercial Use:** allowed/not allowed.
**Special populations:** Ethnicities, gender, pediatric, etc.
**Future use for Methods Development, Aggregate Statistics, Controls**

***Review of ~150 Data Use Limitations Letters at Broad demonstrated that ~90% can be structured with the following ontologies***

***We have formed a DAC to compare automated review of access to traditional mode.***

# Trial of Structured Data Use



Data Access Committee

Data Depositors

Data Use Limitations

Project Request Forms

Data Requestors

1

3

2

1) Can data use letters be structured?

2) Can research purpose be structured?

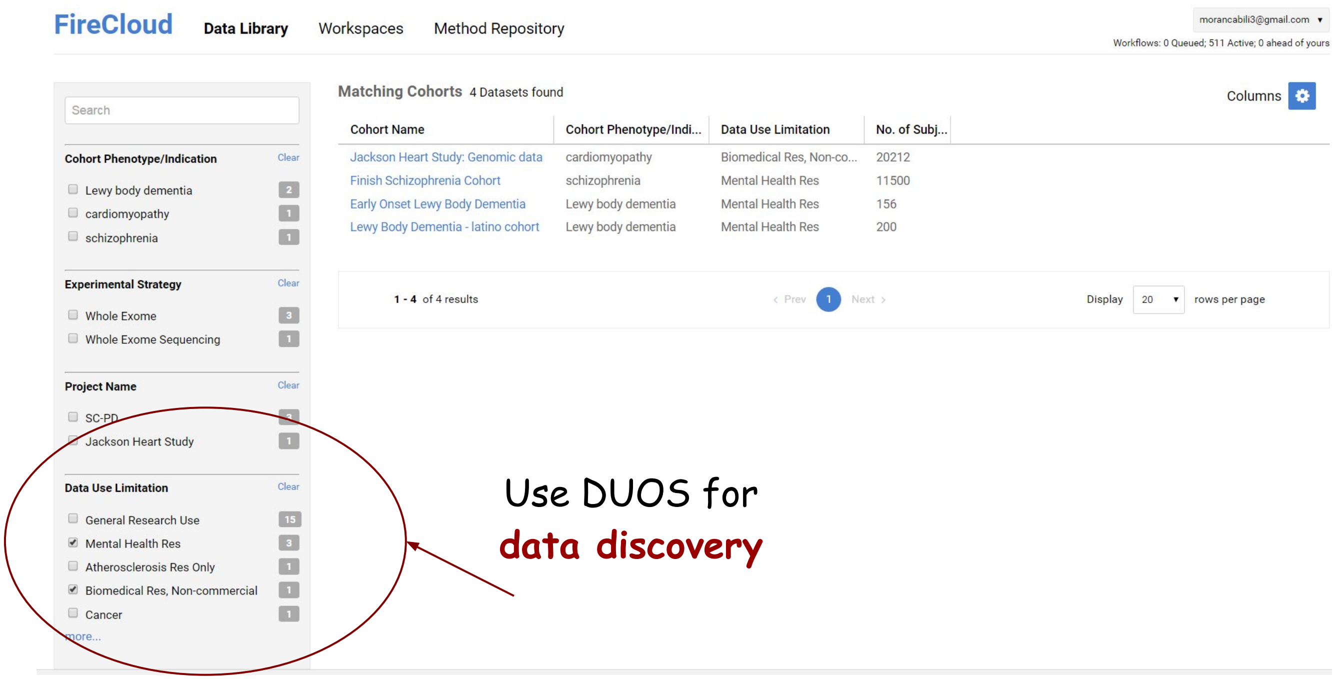
3) Does the DAC agree with the verdict of DUOS?

# What’s next?

1. Updating our ontology to reference DUO
    a. Propose adding new categories to DUO (adopted from ADA-M)

2. DUOS pilot: 3rd batch evaluation of structuring Data Use Limitations

3. Connect to a cloud data repository: Broad’s Data Library on FireCloud



FireCloud   Data Library   Workspaces   Method Repository

morancabili3@gmail.com ▾

Workflows: 0 Queued; 511 Active; 0 ahead of yours

Search

**Cohort Phenotype/Indication** Clear

- □ Lewy body dementia 2
- □ cardiomyopathy 1
- □ schizophrenia 1

**Experimental Strategy** Clear

- □ Whole Exome 3
- □ Whole Exome Sequencing 1

**Project Name** Clear

- □ SC-PD 3
- □ Jackson Heart Study 1

**Data Use Limitation** Clear

- □ General Research Use 15
- ☑ Mental Health Res 3
- □ Atherosclerosis Res Only 1
- ☑ Biomedical Res, Non-commercial 1
- □ Cancer 1

more...

**Matching Cohorts** 4 Datasets found

Columns ⚙

| Cohort Name | Cohort Phenotype/Indi... | Data Use Limitation | No. of Subj... |
|---|---|---|---|
| Jackson Heart Study: Genomic data | cardiomyopathy | Biomedical Res, Non-co... | 20212 |
| Finish Schizophrenia Cohort | schizophrenia | Mental Health Res | 11500 |
| Early Onset Lewy Body Dementia | Lewy body dementia | Mental Health Res | 156 |
| Lewy Body Dementia - latino cohort | Lewy body dementia | Mental Health Res | 200 |

1 - 4 of 4 results

‹ Prev 1 Next ›

Display 20 ▾ rows per page

Use DUOS for **data discovery**

# Acknowledgements

Greg Rushton
Andrea Saltzman
Stacey Donnelly
Anthony Philippakis
Moran Cabili

Pearl O’Rourke
Laura Rodriguez
John Wilbanks

Broad Data Science Platform
DSDE dev team

Stephanie OM Dyke
Anthony Brooks
Melanie Courtot
Dylan Spalding

# DUOS Ontology is Consistent with DUO and ADA-M

| DUOS Restriction code | Dyke et al.  Consent Code | abbreviation | ADA-M |
|---|---|---|---|
| | **Primary Codes: select one** | | |
| No restrictions | no restrictions | NRES | v |
| General research use | general research use and clinical care | GRU(CC) | v |
| Health and Biomedical Research | health/medical/biomedical research and clinical care | HMB(CC) | v |
| Restricted to the study of Disease (ontology ID) | disease-specific research and clinical care | DS-[XX](CC) | v |
| population origins/ancestry research | population origins/ancestry research | POA | v |
| | **Secondary codes: select any** | | |
| Commercial use is prohibited | not-for-profit use only | NPU | v |
| General methods development prohibited | no “general methods” research | NMDS | v |
| Population structure or normal variation studies  prohibited (NPNV)<br>*(*dbGaP question on submission forms)* | N/A | NPNV | v(2 categories) |
| Samples use as controls outside the specified restrictions is prohibited(NCTRL) | NOT MODELED EXPLICITLY;  will be set to True if the following primary categories were selected: disease-specific research and clinical care OR population origins/ancestry research | DS-[XX](CC)<br>OR POA | v |
| Future use of aggregate-level data for general research purposes is prohibited [NAGR] *(*dbGaP question on submission forms)* | N/A | NAGR | N/A |
| limited to gender (M/F/none) | other research-specific restrictions | RS-[GENDER] | v |
| limited to pediatric research | other research-specific restrictions | RS-[PEDIATRIC] | v |
| Sensitive/restricted population; Date | | | |
| N/A | research use only | RUO | v |
| N/A | genetic studies only | GSO | v |
| Recontacting Data Subjects (*may* and *must*) | | | v |
| Other term of use (ethics committee req, cloud storage prohibited, geographical restriction) | ethics approval required,geographical restrictions | IRB, GS[XX] | v |

# Problem: Data Use is not Coded!

***Data Use Restrictions**: What are you doing with the data?*

"The donor wants her data used only for non-commercial cancer research"

| | | Satisfies Data Use Restrictions | |
|---|---|---|---|
| | | **Yes** | **No** |
| Appropriate Permissions | **Yes** | **Available** | **No access** |
| | **No** | **No access** | **No access** |

***Permissions**: Who are you?*

"Only consortium members can READ this data until it is published."

***Main Question: Can Data Use Restrictions be made machine-readable?***